Descripción de la tarea

La empresa Construcciones D8 se ha puesto en contacto con la empresa consultora en la que trabajas para que les realicéis un prediseño de lo que sería un sistema Big Data para resolver las siguientes necesidades.

• Hay distintas fuentes externas a su empresa que producen datos interesantes para ellos y les interesaría poder conectarse a ellas para obtenerlos.

• Esas fuentes tienen conjuntos de datos estáticos o que se actualizan anualmente.

• Además hay fuentes internas de la propia empresa que generan datos de forma contínua y hay que ir los obteniendo sobre la marcha.

• La cantidad de datos actualmente es de aproximadamente 500TB, y calculan que se producen otros 100TB nuevos cada año.

• Quieren poder mantener almacenados todos esos datos de modo no se pierdan y además accesibles en todo momento.

• Se realizan transacciones debido a la interacción con clientes en el día a día.

• La junta directiva se reúne una vez al mes y quiere poder acceder a un cuadro de mandos para ver analíticas descriptivas que empleen todos los datos que estuviesen disponibles una semana antes de reunirse. Tales analíticas deben ser interactivas, siendo los directivos capaces de realizar filtrados de información de modo que las gráficas mostradas se actualicen según la información seleccionada.

• Quieren poder decidir a qué clientes ofrecerles ciertas ofertas en función de lo que se sabe de su comportamiento pasado.

Apartado 1: Prediseña un sistema para Big Data

Crea un documento en el que explicas cómo sería el sistema a emplear para resolver las necesidades Big Data del supuesto práctico. Deberás:

• Indicar qué habrá que hacer para ir aumentando la capacidad del clúster según se reciben nuevos datos.

Para controlar el crecimiento anual de 100TB y que el sistema siga siendo eficiente se diseñará un clúster escalable, horizontalmente y verticalmente. Horizontalmente, se agregarán más nodos al clúster de procesamiento y almacenamiento para aumentar la capacidad total; en relación a lo anteriormente mencionado se habilitarán almacenamientos distribuidos usando tecnologías como Apache Spark, permitiendo un procesamiento distribuido de los datos, aprovechando los recursos que se vayan añadiendo conforme el clúster escale.

• Indicar qué capas de la arquitectura Big Data necesitarán estar presentes como mínimo en el sistema a crear.

La arquitectura mínima presente en el sistema serán las siguientes:

- **Capa de Ingestión de Datos:** para obtener datos de distintas fuentes, internas y externas, se podrían optar por las tecnologías de Kafka y Flume para la captura de datos en tiempo real.

- **Capa de almacenamiento:** donde se gestionarán los datos en bruto, tanto estructurados como no estructurados, un ejemplo serían sistemas gestores de base de datos NoSQL como Cassandra.

- **Capa de procesamiento:** para procesar datos en tiempo real y por lotes (Batch), siendo necesaria una combinación de procesamiento en paralelo y distribuido, Apache Spark o Hadoop MapReduce son las mejores opciones.

-**Capa de acceso:** los datos procesados deben ser accesibles para consultas OLAP, se puede implementar mediante SQL-on-Hadoop como Hive o Presto.

- **Capa de analítica:** necesaria para crear cuadros de mando interactivos para los directivos, ya sea con tecnologías como Tableau, Power BI o Qlik

• Indicar si alguna parte del sistema necesitará cumplir con las características ACID.

Las transacciones serán la parte crítica donde cumplir las características ACID, ya que la empresa interactúa diariamente con clientes y realiza estas transacciones, para garantizar la integridad funcional serán necesarias bases de datos relacionales como MySQL o PostgreSQL además de poder emplearse en conjunto con NoSQL como MongoDB.

• Indicar si será necesario un subsistema OLTP.

Si, será necesario para gestionar las transacciones diarias, el subsistema debe ser rápido y eficiente permitiendo procesar en tiempo real las interacciones de los clientes. Tanto MySQL y MongoDB, al igual que en las características ACID son sólidas opciones.

• Indicar si será necesario un subsistema OLAP.

Seguido del sistema OLTP también será necesario un OLAP para realizar análisis multidimensionales de los datos almacenados permitiendo el acceso a informes detallados y gráficos interactivos. Amazon Redshift o Google Big Query son los candidatos ideales.

• Indicar si habrá un almacén de datos.

También será necesario un almacén de datos para centralizar los datos históricos y transformados permitiendo el análisis y reportes diferenciados por su eficiencia; herramientas como SnowFlake, Amazon Redshift o Google Big Query son esenciales para obtener respuestas rápidas a consultas complejas y de gran volumen.

• Indicar qué estrategia de procesamiento habrá que emplear para poder crear el cuadro de mandos que quiere la junta directiva.

La estrategia que necesita el cuadro de mandos que requiere la junta directiva es una que combine el procesamiento por lotes y en tiempo real. Su respectiva arquitectura se basaría en Kappa o Lambda; seguidamente el procesamiento por lotes, que se llevará a cabo con Spark y Hadoop, se encargará de procesar grandes volúmenes de datos históricos; mientras que el procesamiento en tiempo real, llevado a cabo por Spark Streaming o Apache Kafka, permitirá obtener datos actualizados para generar informes recientes.

• Indicar si será necesario crear modelos predictivos a partir de los datos.

Será necesario para anticipar comportamientos de los clientes y ofrecerles ofertas personalizadas, por lo tanto se usará machine learning y análisis predictivos basados en los datos históricos de comportamiento de los clientes, librerías como TensorFlow, Pytorch podrán solventar problemas puntuales para mejorar aún más los modelos predictivos.